DOCUMENT RESUME

ED 459 822 IR 058 358

AUTHOR Wacholder, Nina; Evans, David K.; Klavans, Judith L.
TITLE Automatic Identification and Organization of Index Terms for Interactive Browsing.
SPONS AGENCY National Science Foundation, Arlington, VA.
PUB DATE 2001-06-00
NOTE 10p.; In: Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (1st, Roanoke, Virginia, June 24-28, 2001). For entire proceedings, see IR 058 348.
CONTRACT IRI-97-12069; CDA-97-53054
AVAILABLE FROM Association for Computing Machinery, 1515 Broadway, New York NY 10036. Tel: 800-342-6626 (Toll Free); Tel: 212-626-0500; e-mail: acmhelp@acm.org. For full text: http://www1.acm.org/pubs/contents/proceedings/dl/379437/.
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Access to Information; Data Processing; *Indexes; Information Seeking; Information Systems; Natural Language Processing; *Online Searching; Subject Index Terms
IDENTIFIERS Browsing; Interactive Systems

ABSTRACT

The potential of automatically generated indexes for information access has been recognized for several decades, but the quantity of text and the ambiguity of natural language processing have made progress at this task more difficult than was originally foreseen. Recently, a body of work on development of interactive systems to support phrase browsing has begun to emerge. This paper considers two issues related to the use of automatically identified phrases as index terms in a dynamic text browser (DTB), a user-centered system for navigating and browsing index terms: (1) What criteria are useful for assessing the usefulness of automatically identified index terms? (2) Is the quality of the terms identified by automatic indexing such that they provide useful access to document content? The terms this paper focuses on have been identified by LinkIT, a software tool for identifying significant topics in text. Over 90% of the terms identified by LinkIT are coherent and therefore merit inclusion in the dynamic text browser. Terms identified by LinkIT are input to Intell-Index, a prototype DTB that supports interactive navigation of index terms. The distinction between phrasal heads (the most important words in a coherent term) and modifiers serves as the basis for a hierarchical organization of terms. This linguistically motivated structure helps user to efficiently browse and disambiguate terms. The paper conclude that the approach to information access discussed is very promising, and that there is much room for further research. In the meantime, this research is a contribution to the establishment of a solid foundation for assessing the usability of terms in phrase browsing. (Contains 25 references.) (Author/AEF)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 459 822

Automatic Identification and Organization of Index Terms for Interactive Browsing

Nina Wacholder
Columbia University
New York, NY
nina@cs.columbia.edu

David K. Evans
Columbia University
New York, NY
devans@cs.columbia.edu

Judith L. Klavans
Columbia University
New York, NY
klavans@cs.columbia.edu

ABSTRACT

The potential of automatically generated indexes for information access has been recognized for several decades (e.g., Bush 1945 [2], Edmundson and Wyllys 1961 [4]), but the quantity of text and the ambiguity of natural language processing have made progress at this task more difficult than was originally foreseen. Recently, a body of work on development of interactive systems to support phrase browsing has begun to emerge (e.g., Anick and Vaithyanathan 1997 [1], Gutwin et al. [10], Nevill-Manning et al. 1997 [17], Godby and Reighart 1998 [9]). In this paper, we consider two issues related to the use of automatically identified phrases as index terms in a dynamic text browser (DTB), a user-centered system for navigating and browsing index terms: 1) What criteria are useful for assessing the usefulness of automatically identified index terms? and 2) Is the quality of the terms identified by automatic indexing such that they provide useful access to document content?

The terms that we focus on have been identified by LinkIT, a software tool for identifying significant topics in text [7]. Over 90% of the terms identified by LinkIT are coherent and therefore merit inclusion in the dynamic text browser. Terms identified by LinkIT are input to Intell-Index, a prototype DTB that supports interactive navigation of index terms. The distinction between phrasal heads (the most important words in a coherent term) and modifiers serves as the basis for a hierarchical organization of terms. This linguistically motivated structure helps users to efficiently browsing and disambiguate terms. We conclude that the approach to information access discussed in this paper is very promising, and also that there is much room for further research. In the meantime, this research is a contribution to the establishment of a solid foundation for assessing the usability of terms in phrase browsing applications.

Keywords

Indexing, phrases, natural language processing, browsing, genre.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

1. OVERVIEW

Indexes are useful for information seekers because they:

- support browsing, a basic mode of human information seeking [17].
- provide information seekers with a valid list of terms, instead of requiring users to invent the terms on their own. Identifying index terms has been shown to be one of the hardest parts of the search process, e.g., [8].
- are organized in ways that bring related information together [16].

But indexes are not generally available for digital libraries. The manual creation of an index is a time consuming task that requires a considerable investment of human intelligence [16]. Individuals and institutions simply do not have the resources to create expert indexes for digital resources.

However, automatically generated indexes have been legitimately criticized by information professionals such as Mulvany 1994 [16]. Indexes created by computer systems are different than those compiled by human beings. A certain number of automatically identified index terms inevitably contain errors that look downright foolish to human eyes. Indexes consisting of automatically identified terms have been criticized on the grounds that they constitute indiscriminate lists, rather than synthesized and structured representation of content. And because computer systems do not understand the terms they extract, they cannot record terms with the consistency expected of indexes created by human beings.

Nevertheless, the research approach that we take in this paper emphasizes fully automatic identification and organization of index terms that actually occur in the text. We have adopted this approach for several reasons:

1. **Human indexers simply cannot keep up with the volume of new text being produced.** This is a particularly pressing problem for publications such as daily newspapers which are under particular pressure to rapidly create useful indexes for large amounts of text.
2. **New names and terms are constantly being invented and/or published.** For example, new companies are formed (e.g., *Verizon Communications Inc.*); people's names appear in the news for the first time (e.g., it is unlikely that *Elian Gonzalez'* name was in a newspaper before November 25, 1999); and new product names are constantly being invented (e.g., *Handspring's Visor PDA*). These terms frequently appear in print some time before they appear in an authoritative reference source.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

D. Cotton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

IR058358

1

BEST COPY AVAILABLE

2

3. **Manually created external resources are not available for every corpus.** Systems that fundamentally depend on manually created resources such as controlled vocabularies, semantic ontologies, or manually annotated text usually cannot be readily adopted to corpora for which these resources do not exist.

4. **Differing indexing standards across agencies and organizations makes reconciliation of indexes a difficult and time consuming task.** The difficulty of reconciliation is exacerbated when indexes are prepared by different organizations for different user groups, corpora and domains (Hert et al. 2000 [11]). Under some circumstances, it may be preferable to have one large automatically generated index than none at all.

5. **Automatically identified index terms are useful in other digital library applications.** Index term lists are essential for browsing, but can also form data as input to other applications such as information retrieval, summarization and classification [25], [1].

Given these information needs, our goal is develop techniques for identifying and organizing index terms that reduce the number of terms that users need to browse and simultaneously maximize the informativeness of each term.

In this paper, we describe a method for identifying index terms for use in a dynamic text browser (DTB). We have implemented a prototype DTB called Intell-Index which supports interactive navigation of index terms, with hyperlinks to the views of phrases in context and full-text documents. The input to Intell-Index consists of noun phrases identified by LinkIT, a software tool for identifying significant topics in domain independent text. (We describe this software in more detail in Section 2.)

However, little work has been done on the question of what constitutes useful index terms (Milstead 1994 [15], Hert et al. 2000 [11]). In order to move toward our goal, we have therefore found it necessary to identify properties of index terms that affect their usefulness in an electronic browsing environment. To assess the quality of the index terms, we consider three criteria especially pertinent to automatically identified index terms: **coherence, thoroughness of coverage of document content,** and a combined metric of quality and coverage that we call **usefulness**.

- **Coherence:** Because computer systems are unable to identify terms with human reliability or consistency, they inevitably generate some number of junk terms that humans readily recognize as incoherent. We consider a very basic question: are automatically identified terms sufficiently coherent to be useful as access points to document content. To answer this question for the LinkIT output, we randomly selected .025% of the terms identified in a 250MB corpus and evaluated them with respect to their coherence. Our study showed that over 90% of the terms are coherent. Cowie and Lehnert 1996 [3] observe that 90% precision in information extraction is probably satisfactory for every day use of results; this assessment is relevant here because the terms are processed by people, who can fairly readily ignore the junk if they expect to encounter it.

- **Thoroughness of coverage of document content:** Because computer systems are more thorough and less discriminating, they typically identify many more terms than a human indexer would for the same amount of material. For example, LinkIT identifies about 500,000 non-unique terms for 12.27 MB of text. We address the issue of quantity by considering the number of terms that LinkIT identifies, as related to size of the original text from which they were extracted. This provides a basis for future comparison of the number of terms identified in different corpora and by different techniques.

- **Usefulness of index terms:** Techniques for automatically identifying index terms abound. In order to make a preliminary assessment of the usefulness of index terms identified by LinkIT, we performed an experiment to measure user's perceptions of the usefulness of index terms. We presented users with lists of index terms identified by three domain-independent techniques and with the newspaper articles from which the terms had been extracted (Wacholder et al. 2000 [22]). In terms of quality alone, our results show that the technical term extraction method of Justeson and Katz 1995 [14] receives the highest rating. However, in terms of a combined quality and coverage metric, the Head Sorting method, described in Wacholder 1998 [21] used by LinkIT outperforms both other techniques.

Although the terms identified by LinkIT are the primary focus of our analysis, these criteria can be applied systematically to terms identified by other techniques. Techniques for identifying NPs abound and they are difficult to compare (Evans et al. 2000 [7]). With this work, we seek to establish a foundation for further in-depth analysis of the usability of automatically identified terms which will include, *inter alia*, the observation of subjects performing information seeking tasks using index terms generated by different techniques.

We discuss of term coherence, usefulness and thoroughness of coverage of document content in Section 3. But before we turn to this discussion, we describe the technology that we have developed to identify, display, and structure terms.

2. Technology

2.1 Towards a DTB

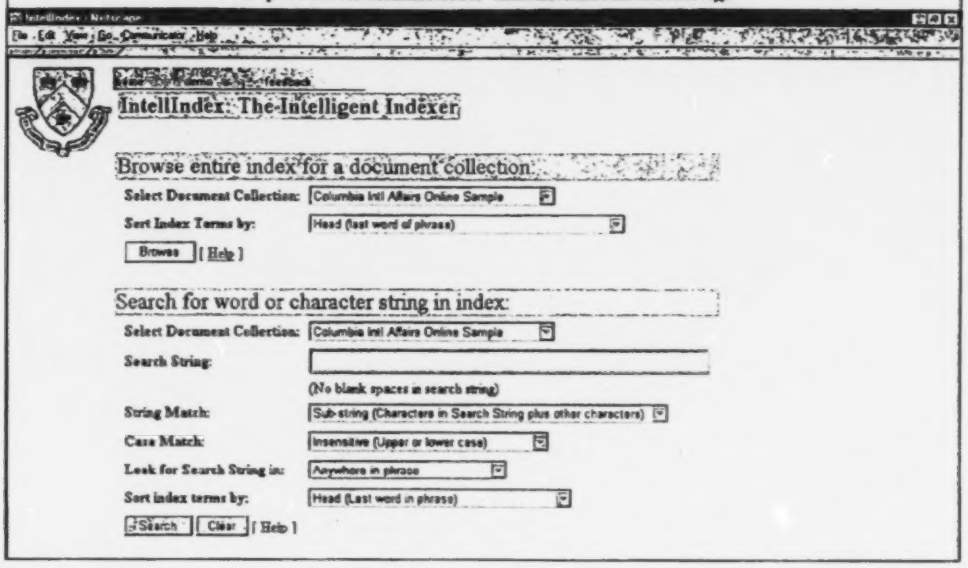
We have designed a domain-independent system called LinkIT, which uses the head sorting method to identify candidate index terms in full-text documents (Wacholder 1998 [21]; Evans 1998[6]). Using a finite state machine compiled from a regular expression grammar, LinkIT parses text which has been automatically tagged with grammatical part-of-speech a finite state machine. LinkIT can process approximately 4.11 MB tagged text per second [7], [6].

The expressions identified by LinkIT are noun phrases (NPs), coherent units whose head (most important word syntactically and semantically) is a noun. For example, *filter* is the head of the NPs *coffee filter*, *oil filter*, *smut filter*, and *water filters at warehouse prices*.

At the present time LinkIT identifies a subset of NPs that occur in a document, simplex NPs. A complex NP *a form of cancer-causing asbestos* actually includes two simplex NPs, *a form* and *cancer-causing asbestos*. A system that lists only complex NPs would list only one term, a system that lists both simplex and complex NPs would list all three phrases, and a system that identifies only simplex NPs would list two. LinkIT identifies Simplex NPs rather than complex ones for a practical reason: Simplex NPs can be identified more reliably because they are structurally more simple. Compared to simplex NPs, the boundaries of complex NPs (e.g., *symptoms that*

Figure 1: Intell-Index opening screen

< http://www.cs.columbia.edu/~nina/IntellIndex/indexer.cgi>


IntellIndex: The Intelligent Indexer
Browse entire index for a document collection:
Select Document Collection:
Columbia Intl Affairs Online Sample
Sort Index Terms by:
Head (last word of phrase)
Browse
[Help]
Search for word or character string in index:
Select Document Collection:
Columbia Intl Affairs Online Sample
Search String:
(No blank spaces in search string)
String Match:
Sub-string (Characters in Search String plus other characters)
Case Match:
Insensitive (Upper or lower case)
Look for Search String in:
Anywhere in phrase
Sort index terms by:
Head (Last word in phrase)
Search
Clear
[Help]

crop up decades later) are difficult to identify. The head is also more difficult to identify: for example there are cases, such as *group of children* where the syntactic head (*group*) is distinct from the semantic head (*children*). Complex NPs can be difficult for people to interpret, especially out of context. For example, the expression *information about medicine for babies* is ambiguous: in [[information about medicine] [for infants]], the information is for infants; in [information about [medicine for infants]], the medicine is for infants. The decision to include only simplex NPs in the DTB has important implications for the number of index terms included in the DTB, as discussed below.

Finally, LinkIT sorts the NPs by head, and ranks them in terms of their significance based on head frequency. The intuitive justification for sorting SNPs by head is based on the fundamental linguistic distinction between head and modifier: in general, a head makes a greater contribution to the syntax and semantics of a phrase than does a modifier. This linguistic insight can be extended to the document level. If, as a practical matter, it is necessary to rank the contribution to a whole document made by the sequence of words constituting a documentre, the head should be ranked more highly than other words in the phrase. This distinction is important in linguistic theory; for example, Jackendoff 1977 [13] discusses the relationship of heads and modifiers in phrase structure. It is also important in NLP, where, for example, Strzalkowski 1997 [19] and Evans and Zhai 1996 [5] have used the distinction between heads and modifiers to add query terms to information retrieval systems.

Powerful corpus processing techniques have been developed to measure deviance from an average occurrence or co-occurrence in the corpus. In this paper we chose to evaluate methods that depend only on document-internal data, independent of corpus, domain or genre. We therefore did not use, for example, tf*idf, the purely statistical technique that is the used by most information retrieval systems, or Smadja 1993 [17], a hybrid statistical and symbolic technique for identifying collocations.

We have incorporated LinkIT output into a prototype DTB called Intell-Index.
(http://www.cs.columbia.edu/~nina/IntellIndex/indexer.cgi). Figure 1 above shows the Intell-Index opening screen. The user selects the collection to be browsed and then may browse the entire set of index terms identified by LinkIT. (Note that this "collection" could also arise from the results of a search.) Alternatively, the user may enter a term, and specify criteria to select a subset of terms that will be returned (e.g. heads only, or modifiers and heads). This gives the user better control over the results so that browsing is more effective.

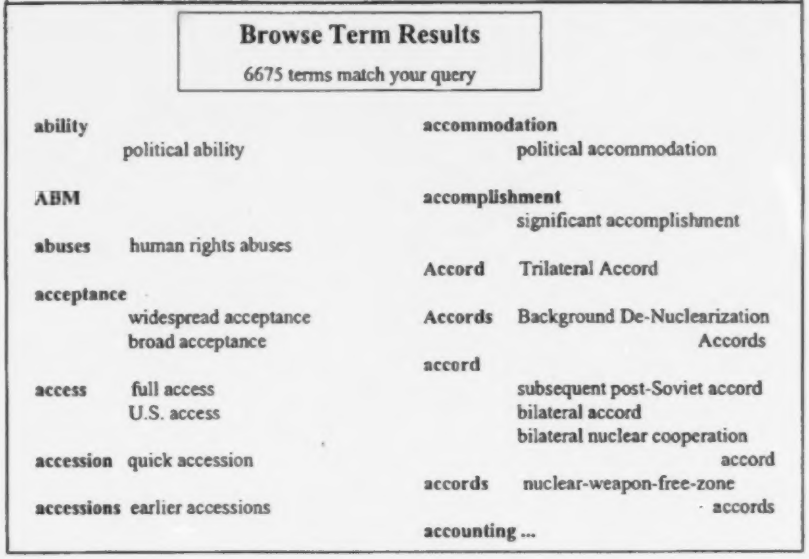
Figure 2 on p.4 shows the beginning of the alphabetized browsing results for the specified corpus. As the user browses the terms returned by Intell-Index, she may choose to view a list of the contexts in which the terms are used; these contexts are sorted by document and ranked by normalized frequency in the document. This view is called index term in context (ITIC) based on its relationship to a simpler version, i.e. keyword in context (KWIC). If the information seeker decides that the list of ITICs is promising, she may view the entire document, or browse another view of the data.

At the present time, the DTB uses only Simplex NPs. However, LinkIT collects information for conflating simplex NPs into complex ones; this will be added to Intell-Index at a later date.

2.2 The head sorting technique

A key advantage of using Simplex NPs rather than Complex ones is that, at least in English, the last word of the NP is reliably the head (Wacholder 1998 [21]). Repetition of heads of phrases in a

Figure 2: Browse term results



document indicates that the head represents an important concept in the document. As a result, no additional information other than that extracted from the document is required to sort the NPs by head.

Information about frequency with which nouns are used as heads in documents can be used to provide the users with useful views of the content of a single document or a collection of documents. Table 1 shows the topics are identified as most important in a single article using the head sorting technique (*Wall Street Journal 1988* [23]). Heads of terms are italicized.

asbestos *workers*
cancer-causing *asbestos*
cigarette *filters*
researcher(s)
asbestos *fiber*
crocidolite
paper *factory*

Table 1: Most significant terms in document

For example the list of phrases (which includes heads that occur above a frequency cutoff of 3 in this document, with content-bearing modifiers, if any) is a list of important concepts representative of the entire document.

Another view of the phrases enabled by head sorting is obtained by linking NPs in a document with the same head. A single word NP can be quite ambiguous, especially if it is a frequently-occurring noun like *worker*, *state*, or *act*. NPs grouped by head are likely to refer to the same concept, if not always to the same entity (Yarowsky 1993 [24]), and therefore convey the primary sense of the head as used in the text. For example, in the sentence "Those workers got a pay raise but the other workers did not", the same sense of *worker* is used in both NPs even though two different sets of workers are referred to. Table 2 shows how the word *workers* is used as the head of a NP in four different Wall Street Journal articles from the Penn Treebank; determiners such as *a* and *some* have been removed.

workers ... asbestos workers (wsj 0003)
workers ... private sector workers ... private sector hospital workers ... nonunion workers...private sector union workers (wsj 0319)
workers ... private sector workers ... United Steelworkers (wsj 0592)
workers ... United Auto Workers ... hourly production and maintenance workers (wsj0492)

Table 2: Comparison of uses of *worker* as head of NPs across articles

This view distinguishes the type of *worker* referred to in the different articles, thereby providing information that helps rule in certain articles as possibilities and eliminate others. This is because the list of complete uses of the head *worker* provides explicit positive and implicit negative evidence about kinds of workers discussed in the article. For example, since the list for wsj_0003 includes only *workers* and *asbestos workers*, the user can infer that hospital workers or union workers are probably not referred to in this document.

Term context can also be useful if terms are presented in document order. For example, the index terms in Table 3 were extracted

automatically by the LinkIT system as part of the process of identification of all NPs in a document (Evans 1998 [6]; Evans et al. 2000[7]).

A form
asbestos
Kent cigarette filters
a high percentage
cancer deaths
a group
workers
30 years
researchers

Table 3: Topics, in document order, extracted from first sentence of *Wall Street Journal* article

For most people, it is not difficult to guess that this list of terms has been extracted from a discussion about deaths from cancer in workers exposed to asbestos. The information seeker is able to apply common sense and general knowledge of the world to interpret the terms and their possible relation to each other. At least for a short document, a complete list of terms extracted from a document in order can relatively easily be browsed in order to get a sense of the topics discussed in a single document.

In the next section we assess, qualitatively and quantitatively, the usability of automatically indexed terms identified by LinkIT. The focus of this discussion is Simplex NPs; in future work, we will discuss techniques for extending the informativeness of Simplex NPs.

3. Assessment of automatically identified index terms

The problem of how to determine what index terms merit inclusion in a DTB is a difficult one. The standard information retrieval metrics of precision and recall do not apply to this task because indexes are designed to satisfy multiple information needs. In information retrieval, precision is calculated by determining how many retrieved documents satisfy a specific information need. But indexes by design include index terms that are relevant to a variety of information needs. To apply the recall metric to index terms, we would calculate the proportion of good index terms correctly identified by a system relative to the list of all possible good index terms. But we do not know what the list of all possible good index terms should look like. Even comparing an automatically generated list to a human generated list is difficult because human indexers add index entries that do not appear in the text; this would bias the evaluation against an index that only includes terms that actually occur in the text. We have therefore identified three basic criteria that affect usability of index terms: coherence of terms, thoroughness of coverage of document content, and usefulness.

3.1 Coherence

For index terms to be useful, they must be coherent. This criterion is particular important because any list of automatically identified index terms inevitably includes some junk. An index with less junk terms is clearly superior to one with more junk.

To assess the coherence of automatically identified index terms, 583 index terms (.025% of the total number of terms identified) were randomly extracted from the 250 MB corpus and alphabetized. Each term was assigned one of three ratings:

- **coherent** – a term is both coherent and an NP. Coherent terms make sense as a distinct unit, even out of context. Examples of coherent terms identified by LinkIT are *sudden current shifts*, *Governor Dukakis*, *terminal-to-host connectivity* and *researchers*.
- **incoherent** – a term is neither a NP nor coherent. Examples of incoherent terms identified by LinkIT are *uncertainty is*, *x ix limit*, and *heated potato then shot*. Most of these problems result from idiosyncratic or non-standard text formatting. Another source of errors is the part-of-speech tagger; for example, if it erroneously identifies a verb as a noun (as in the example *uncertainty is*), the resulting term is incoherent.
- **intermediate** – any term that does not clearly belong in the coherent or incoherent categories. Typically they consist of one or more good NPs, along with some junk. In general, they are enough like NPs that in some ways they fit patterns of the component NPs. One example is *up Microsoft Windows*, which would be a coherent term if it did not include *up*. We include this term because the term is coherent enough to justify inclusion in a list of references to Windows or Microsoft. Another example is *th newsroom*, where *th* is presumably a typographical error for *the*. There are a higher percentage of intermediate terms among proper names than the other two categories; this is because LinkIT has difficulty of deciding where one proper name ends and the next one begins, as in *General Electric Co. MUNICIPALS Forest Reserve District*.

Table 4 shows the ratings by type of term and overall. The percentage of useless terms is 6.5%. This is well under 10%, which puts our results in the realm of being suitable for everyday use according to the Cowie and Lehnert metric mentioned in Section 1.

	Total	Cohe-rent	Inter-mediate	Inco-herent
Number of words	574	475	62	37
% of total words	100%	82.8%	10.9%	6.5%

Table 4: Coherence of index terms

This study demonstrates that automatically identified terms like those identified by LinkIT are of sufficient quality to be useful in browsing applications.

3.2 Usefulness of index terms

In order to make a preliminary determination of whether the terms identified by LinkIT as likely to be useful index terms, we used a standard qualitative ranking technique to compare the usefulness of terms identified by three domain-independent techniques methods for identifying index terms (Wacholder et al. 2000 [22]):

- **Keywords** are terms identified by counting frequency of stemmed words in a document.
- **Technical terms** are noun phrases (NPs) or subparts of NPs repeated more than twice in a document (Justeson and Katz 1995 [14]);

- **Head sorted NPs** are identified by a method in which simplex noun phrases (as defined below) are sorted by head and then ranked in decreasing order of frequency (Wacholder 1998 [21]).

Table 5 shows examples of the index terms identified by the different techniques. All technical terms are included; a sample of the terms identified by the other two techniques are included.

Keywords	Head sorted NPs	Technical terms
asbestos/asbestosis	workers	cancer deaths
worker/workers /worked	asbestos workers	lung cancer
cancer	160 workers	kent cigarette
death	cancer	dr. talcott
make	lung cancer	cigarette filter
lorillard	asbestos	u.s.
fiber	cancer causing asbestos	
dr.	lung cancer deaths	
...	...	

Table 5: Examples of terms, by technique for one document

To compare the index terms, we presented subjects with an article from the *Wall Street Journal* [23] and a list of terms and asked them to answer the following general question: "Would this term be useful in an electronic index for this article?" Terms were rated on a scale of 1 to 5, where 1 indicates a high quality term that should definitely be included in the index and 5 indicates a junk term that definitely should not be included. For example, the phrase *court-approved affirmative action plans* received an average rating of 1 meaning that it was ranked as definitely useful; the keyword *affirmative* received an average rating of 3.75, meaning that it was less useful; and the keyword *action* received an average ranking of 4.5, meaning that it was not useful. Table 6 shows the results, averaged over three articles.

	Keywords	Head Sorted NPs	Technical Terms
Average ranking	3.27	2.89	1.79

Table 6 : Average rating of types of index terms

Of the three lists of index terms, technical termss received the highest ratings for all three documents—an average of 1.79 on the scale of 1 to 5, with 1 being the best rating. The head sorted NPs came in second, with an average of 2.89, and keywords came in last with an average of 3.27.

However, it should be noted that the average ranking of terms conceals the fact that the number of technical terms is much lower than the other two types of terms. In contrast, Table 7, which shows the total number of terms rated at or below specified rankings, allows us to measure quality and coverage. (1 is the highest rating; 5 is the lowest.)

	Number of terms ranked at or better than			
Method	2	3	4	5
Keyword	27	75	124	166
Head sorted NPs	41	96	132	160
Technical Terms	15	21	21	21

Table 7: Running total of terms identified at or below a specified rank

This result is consistent with our observation that the technical term method identifies the highest quality terms, but there are very few of them: an average of 7 per 500 words compared to over 50 for head sorted NPs and for keywords. Therefore there is a need for additional high quality terms. The list of head sorted NPs received a higher average rating than did the list of keywords, as shown in Figure 2. This confirms our expectation that phrases containing more content-bearing modifiers would be perceived as more useful index terms than would single word phrases consisting only of heads.

3.3 Thoroughness of coverage of document content

Thoroughness of coverage of document content is a standard criterion for evaluation of traditional indexes [11]. In order to establish an initial measure of thoroughness, we evaluate number of terms identified relative to the size of the text.

Table 8 shows the relationship between document size in words and number of NPs per document. For example, for the AP corpus, an average document of 476 words typically has about 127 non-unique NPs associated with it. In other words, a user who wanted to view the context in which each NP occurred would have to look at 127 contexts. (To allow for differences across corpora, we report on overall statistics and per corpus statistics as appropriate.)

Corpus	Avg. Doc Size	Avg. number of NPs/doc
AP	2.99K (476 words)	127
FR	7.70K (1175 words)	338
WSJ	3.23K (487 words)	132
ZIFF	2.96K (461 words)	129

Table 8: NPs per document

The numbers in Table 8 are important because they vary depending on the technique used to identify NPs. A human indexer readily chooses whichever type of phrase is appropriate for the content, but natural language processing systems cannot do this reliably. Because of the ambiguity of natural language, it is much easier to identify the

boundaries of simplex noun than complex ones [21], as discussed above in Section 2 above.

The option of including both complex and simple forms was adopted by Tolle and Chen 2000 [20]. They identify approximately 140 unique NPs per abstract for 10 medical abstracts. They do not report the average length in words of abstracts, but a reasonable guess is probably about 250 words per abstract. On this calculation, the ratio between the number of NPs and the number of words in the text is .56. In contrast, LinkIT identifies about 130 NPs for documents of approximately 475 words, for a ratio of .27. The index terms represent the content of different units: 140 index terms represents the abstract, which is itself only an abbreviated representation of the document. The 130 terms identified by LinkIT represent the entire text, but our intuition is that it is better to provide coverage of full documents than of abstracts. Experiments to determine which technique is more useful for information seekers are needed.

For four large full-text corpora, we extracted all occurrences of all NPs (duplicates not removed) in each corpus, and then we list the size of the index when duplicate NPs have been removed in Table 9. The numbers in parenthesis are the number of words per document and per corpus for the full-text columns, and the percentage of the full text size for the list of non-unique NPs, and list of unique NPs.

Corpus	Full Text	Non Unique NPs	Unique NPs
AP	12.27 MB (2.0 million words)	7.4 MB (60%)	2.9 MB (23%)
FR	33.88 MB (5.3 million words)	20.7 MB (61%)	5.7 MB (17%)
WSJ	45.59 MB (7.0 million words)	27.3 MB (60%)	10.0 MB (22%)
ZIFF	165.41 MB (26.3 million words)	108.8 MB (66%)	38.7 MB (24%)

Table 9: Corpus Size

The number of NPs reflects the number of occurrences (tokens) of NPs. Interestingly, the percentages are relatively consistent across corpora.

From the point of view of the index, however, the first column in Table 9 represent only a first level reduction in the number of candidate index terms: for browsing and indexing, each term need be listed only once. After duplicates have been removed, approximately 1% of the full text remains for heads, and 22% for NPs. The implications of this are explored in Section 4.

4. Reducing documents to NPs

In this section, we consider how information about NPs and their heads can help facilitate effective browsing by reduce the number of terms that an information seeker needs to look at.

In general, the number of unique NPs increases much faster than the number of unique heads – this can be seen by the fall in the ratio of unique heads to NPs as the corpus size increases.

Corpus	Size in MBs	Unique NPs	Unique Heads	Ratio of Unique Heads to NPs
AP	12	156798	38232	24%
FR	34	281931	56555	20%
WSJ	45	510194	77168	15%
ZIFF	165	1731940	176639	10%
Total	256	2490958	254724	10%

Table 10: Number of unique NPs and heads

Table 10 is interesting for a number of reasons:

1) the variation in ratio of heads to NPs per corpus—this may well reflect the diversity of AP and the FR relative to the WSJ and especially Ziff.

2) the number of heads decreases monotically as the size of the corpus increases. This is because the heads are nouns. No dictionary can list all nouns; this list is constantly growing, but at a slower rate than the possible number of NPs.

In general, the vast majority of heads have two or fewer different possible expansions. There is a small number of heads, however, that contain a large number of expansions. For these heads, we could create a hierarchical index that is only displayed when the user requests further information on the particular head. In the data that we examined, on average the heads had about 6.5 expansions, with a standard deviation of 47.3.

Corp	Max	% <= 2	2 < % < 50	% >= 50	Avg	Std. Dev.
AP	557	72.2%	26.6%	1.2%	4.3	13.63
FR	1303	76.9%	21.3%	1.8%	5.5	26.95
WSJ	5343	69.9%	27.8%	2.3%	7.0	46.65
ZIFF	15877	75.9%	21.6%	2.5%	10.5	102.38

Table 11: Average number of head expansions per corpus

The most frequent head in the Ziff corpus, a computer publication, is *system*.

Additionally, these terms have not been filtered; we may be able to greatly narrow the search space if the user can provide us with further information about the type of terms they are interested in. For example, using simple regular expressions, we are able to roughly categorize the terms that we have found into four categories: NPs, SNPs that look like proper nouns, SNPs that look like acronyms, and SNPs that start with non-alphabetic characters. It is possible to narrow the index to one of these categories, or exclude some of them from the index.

Corpus	# of SNPs	# of Proper Nouns	# of Acronyms	# of non-alphabetic elements
AP	156798	20787 (13.2%)	2526 (1.61%)	12238 (7.8%)
FR	281931	22194 (7.8%)	5082 (1.80%)	44992 (15.95%)
WSJ	510194	44035 (8.6%)	6295 (1.23%)	63686 (12.48%)
ZIFF	1731940	102615 (5.9%)	38460 (2.22%)	193340 (11.16%)
Total	2490958	189631 (7.6%)	45966 (1.84%)	300373 (12.06%)

Table 12: Number of SNPs by category

For example, over all of the corpora, about 10% of the SNPs start with a non-alphabetic character, which we can exclude if the user is searching for a general term. If we know that the user is searching specifically for a person, then we can use the list of proper nouns as index terms, further narrowing the search space to approximately 10% of the possible terms. We regard this technique as an important first step to reducing the number of terms that users must browse in a DTB.

5. CONCLUSION

When we began working on this paper, our goal was simply to assess the quality of the terms automatically identified by LinkIT for use in electronic browsing applications. Through an evaluation of the results of an automatic index term extraction system, we have shown that automatically generated indexes can be useful in a dynamic text-browsing environment such as Intell-Index for enabling access to digital libraries.

We found that natural language processing techniques have reached the point of being able to reliably identify terms that are coherent enough to merit inclusion in a dynamic text browser: over 93% of the index terms extracted for use in the Intell-Index system have been shown to be useful index terms in our study. This number is a baseline; the goal for us and others should be to improve these numbers.

We have also demonstrated how sorting of index terms by head makes it easier to browse index terms. The possibilities for additional sorting and filtering index terms are multiple, and our work suggests that these possibilities are worthy of exploration. Our results have implications for our own work and also for research results with regard to phrase browsers referred to in Section 1.

As we conducted this work, we discovered that there are many unanswered questions about the usability of index terms. In spite of a long history of indexes as an information access tool, there has been relatively little research on indexing usability, an especially important topic vis a vis automatically generated indexes [11][15].

Among them are the following:

1. What properties determine the usability of index terms?
2. What techniques for automatically identifying index terms produce the most useful index terms?
3. How is the usefulness of index terms affected by the browsing environment, the domain of the document, and the user's expertise?
4. From the point of view of representation of document content, what is the optimal relationship between number of index terms and document size?
5. What number of terms can information seekers readily browse? Do these numbers vary depending on the skill and domain knowledge of the user?

Because of the need to develop new methods to improve access to digital libraries, answering questions about index usability is a research priority in the digital library field. This paper makes two contributions: description of a linguistically motivated method for identifying and browsing index terms and establishment of fundamental criteria for measuring the usability of terms in phrase browsing applications.

6. ACKNOWLEDGMENTS

This work has been supported under NSF Grant IRI-97-12069, "Automatic Identification of Significant Topics in Domain Independent Full Text Analysis", PI's: Judith L. Klavans and Nina Wacholder and NSF Grant CDA-97-53054 "Computationally Tractable Methods for Document Analysis", PI: Nina Wacholder.

7. References

[1] Anick, Peter and Shivakumar Vaithyanathan (1997) "Exploiting clustering and phrases for context-based information retrieval", *Proc. of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '97), pp.314-323.

[2] Bush, Vannevar (1945) "As we may think," *Atlantic Monthly*. Available from http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm

[3] Cowie, Jim and Wendy Lehnert (1996) "Information extraction", *Communications of the ACM*, 39(1):80-91.

[4] Edmundson, H.P. and Wyllys, W. (1961) "Automatic abstracting and indexing--survey and recommendations", *Communications of the ACM*, 4:226-234.

[5] Evans, David A. and Chengxiang Zhai (1996) "Noun-phrase analysis in unrestricted text for information retrieval", *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics*, pp.17-24. 24-27 June 1996, University of California, Santa Cruz, California, Morgan Kaufmann Publishers.

[6] Evans, David K. (1998) LinkIT Documentation, Columbia University Department of Computer Science Report. Available at <http://www.cs.columbia.edu/~devans/papers/LinkITTechDoc/>

[7] Evans, David K., Klavans, Judith, and Wacholder, Nina (2000) "Document processing with LinkIT", *Proc. of the RIAO Conference*, Paris, France.

[8] Furnas, George, Thomas K. Landauer, Louis Gomez and Susan Dumais (1987) "The vocabulary problem in human-system communication", *Communications of the ACM* 30:964-971.

[9] Godby, Carol Jean and Ray Reighart (1998) "Using machine-readable text as a source of novel vocabulary to update the Dewey Decimal Classification", presented at the SIG-CR Workshop, ASIS, < http://orc.rsch.oclc.org:5061/papers/sigcr98.html >.

[10] Gutwin, Carl, Gordon Paynter, Ian Witten, Craig Nevill-Manning and Eibe Franke (1999) "Improving browsing in digital libraries with keyphrase indexes", *Decision Support Systems* 27(1-2):81-104.

[11] Hert, Carol A., Elin K. Jacob and Patrick Dawson (2000) "A usability assessment of online indexing structures in the networked environment", *Journal of the American Society for Information Science* 51(11):971-988.

[12] Hodges, Julia, Shiyun Yie, Ray Reighart and Lois Boggess (1996) "An automated system that assists in the generation of document indexes", *Natural Language Engineering* 2(2):137-160.

[13] Jackendoff, Ray, (1977), *X-Bar Syntax: A Study of Phrase Structure*, MIT Press, Cambridge, MA.

[14] Justeson, John S. and Slava M. Katz (1995). "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering* 1(1):9-27.

[15] Milstead, Jessica L. (1994) "Needs for research in indexing", *Journal of the American Society for Information Science*.

[16] Mulvany, Nancy (1993) *Indexing Books*, University of Chicago Press, Chicago, IL.

[17] Nevill-Manning, Craig G., Ian H. Witten and Gordon W. Paynter (1997) "Browsing in digital libraries: a phrase based approach", *Proc. of the DL97*, Association of Computing Machinery Digital Libraries Conference, 230-236.

[18] Smadja, Frank, McKeown, Kathy, and Vasileios Hatzivassiloglou, V. (1996). "Translating collocations for bilingual lexicons: A statistical approach". *Computational Linguistics* 22 (1):1-38.

[19] Strzalkowski, Tomek. 1997. "Building Effective Queries in Natural Language Information Retrieval." *Proc. of the 5th Applied Natural Language Conference* (ANLP-97), Washington, DC. pp. 299-306.

[20] Tolle, Kristin M. and Hsinchun Chen (2000) "Comparing noun phrasing techniques for use with medical digital library tools", *Journal of the American Society of Information Science* 51(4):352-370.

[21] Wacholder, Nina (1998) "Simplex noun phrases clustered by head: a method for identifying significant topics in a document", *Proc. of Workshop on the Computational Treatment of Nominals*, edited by Federica Busa, Inderjeet Mani and Patrick Saint-Dizier, pp.70-79. COLING-ACL, October 16, 1998, Montreal.

[22] Wacholder, Nina, David Kirk Evans, Judith L. Klavans (2000) "Evaluation of automatically identified index terms for browsing electronic documents", *Proc. of the Applied Natural Language Processing and North American Chapter of the Association for Computational Linguistics (ANLP-NAACL) 2000*. Seattle, Washington, pp. 302-307.

[23] Wall Street Journal (1988) Available from Penn Treebank, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

[24] Yarowsky, David (1993) "One sense per collocation", *Proc. of the ARPA Human Language Technology Workshop*, Princeton, NJ, pp.266-271.

[25] Zhou, Joe (1999) "Phrasal terms in real-world applications". In *Natural Language Information Retrieval*, edited by Tomek Strazalowski, Kluwer Academic Publishers, Boston, pp.215-259.